MORPHOLOGICAL FILTERS IN FLOODPLAIN FOR DEM-EXTRACTED DATA – USING MINIMUM BOUNDING CIRCLE & YOUDEN INDEX

Peng Jin

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in the Department of Geography,
Indiana University
June 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Master's Thesis Committee

Daniel P. Johnson Ph.D., Chair

Frederick L. Bein Ph.D.

Vijay O Lulla Ph.D.

ACKNOWLEDGMENTS

This research was supported by IUPUI – POLIS Center and IUPUI – Department of Geography. I would like to show my gratitude to the POLIS Center for sharing their data with us. I am sincerely thankful to my colleague Jeffery Ashby for using his classification methods to produce our research data. I would also like to take this opportunity to express my sincere gratitude to my advisor Daniel P. Johnson Ph.D. (Chair) for his continued support throughout my research, his patience, enthusiasm, and immense knowledge. I am also immensely grateful to my thesis committee: Frederick L. Bein Ph.D. and Vijay O Lulla Ph.D. for providing their insight and expertise that greatly assisted my research.

Peng Jin

MORPHOLOGICAL FILTERS ON DEM-EXTRACTED DATA – USING MINIMUM BOUNDING CIRCLE & YOUDEN INDEX

Floods are one of the worst disasters in the United States. Each year, the government allocates a tremendous amount of manpower and money on flood prevention initiatives. As the first defense line, levees provide protection from temporary flooding (Makhdoom, 2013). These embankments are broadly classified according to the areas they protect, which could either be urban or agricultural levees within floodplains. In the U.S., most of the levees are handled by government agencies such as the U.S. Army Corps of Engineers and the Federal Emergency Management Services. On the other hand, non-levee embankments created by individual farmers (Olson & Morton, 2013) or naturally formed levee-like structures may not be in the government database. The initial purpose of this research was to assist Polis center on the “Mapping of Non-Levee Embankments in the Indiana” project. The non-levee embankments are not certified or engineered levee-like structures. They, therefore, impose lateral constraints on flood flows, reducing the floodplain storage capacity and increasing the flood velocity. These non-levee embankments can cause stream erosion and downstream flooding. Therefore, it is important to know the locations of these features. The first part of the proposed method adapted the Empirical Bayesian theorem and the low pass filter techniques to extract elevated linear features from LiDAR elevation data – Digital Elevation Model (DEM). The second part of the proposed methods combined the Minimum Bounding Circle (MBC) method and the Youden Index to locate the optimal threshold value that can be used to determine whether the extracted features are levee-like structures. The focus of this study is not only limited to artificial levee-like structures, but also takes the natural levees, or any potential levee-like structures into account because this study assumes all embankments play important roles during flood events.

Daniel P. Johnson Ph.D

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

DEM	Digital Elevation Model
DTM	Digital Terrian Model
LiDAR	Light/Laser Detection and Ranging
MBC	Minimum Bounding Circle
RCC	Related Circumscribing Circle
ROC	Receiver Operating Characteristic
FEMA	Federal Emergency Management Agency
USACE	U.S. Army Corps of Engineers

INTRODUCTION

Flood impact is one of the most significant disasters in the United States. Flood damage includes a wide range of harmful effects on humans, their health, and their properties (e.g., damaging houses, cars, etc.), on public infrastructure (e.g., flooding roads, breaking dykes, etc.), cultural heritage, ecological systems (e.g., causing pollution), industrial production (e.g., loss of production due to destroyed facility) and the competitive strength of the affected economy (e.g., lacking of supply) (Messner & Meyer, 2006). The Federal Emergency Management Agency (FEMA) and the U.S. Army Corps of Engineers (USACE) have spent a tremendous amount of money on flood control. Therefore, flood control is important as it reduces or prevents damages during the flood. The prevention methods include structural flood control measures, such as the construction of dams, river dikes or flood forecasting and warning, flood hazard and risk management, public participation and institutional arrangement among others (Tingsanchali, 2012).

Levee

As the most common flood defense system, levees play a very significant role during floods. They are defined as raised structures whose primary purpose is to provide protection against floods (Courtesy Beeldbank VenW.nl, 2013; Makhdoom, 2013). These embankments are usually several meters higher than the floodplain and close to the river channels (Bailey, 2007). These morphological features are generally long linear structures (Courtesy Beeldbank VenW.nl, 2013; Steinfeld, Kingsford, & Laffan, 2013). They are also known as flood defense, embankments, digues, floodwalls or dikes (Courtesy Beeldbank VenW.nl, 2013; Fuhrman, 2000; Simm et al., 2012).

There are generally 2 categories of levees, a natural levee, and an artificial levee. Artificial levees are usually made by piling soil, sand, or rocks on a cleared level surface (Figure 1). They also can be built from wood, metal, plastic, or even concrete. Their main purpose is to prevent flooding of the river onto the adjoining land (Bailey, 2007). Natural levees are formed from a long period of flood deposits that form sinuous ridges of sediments along river channels or within floodplains

(Hudson, 2005). Natural levees are very common features of alluvial river systems (Adams, Slingerland, & Smith, 2004). There are 2 types of artificial levees. The types are classified by location: urban levees and agricultural levees. Urban levees are used to prevent flooding in community areas, such as industrial, residential, and commercial areas. Agricultural levees provide protection from flooding in farm lands.

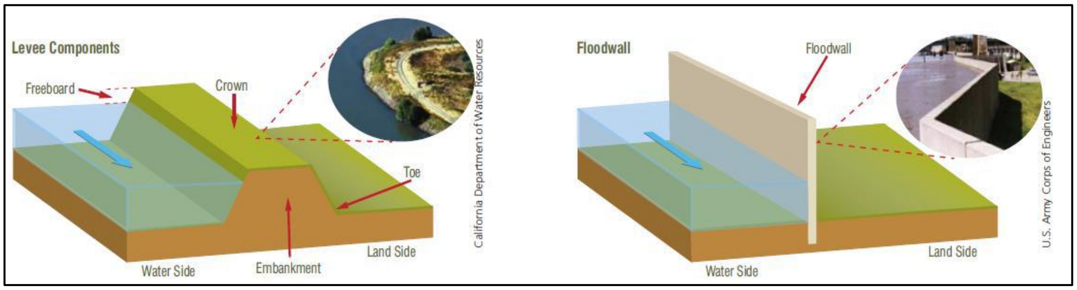


***Figure 1** Natural Levees (left), Artificial Levees (right) (Makhdoom, 2013).*

Levees are further classified according to their use. In this method, there are 5 categories: mainline and tributary levees, ring levees, sub-levees, setback levees, and spur levees. Mainline and tributary levees are generally parallel to the channels. Ring levees completely encircle a small cluster of buildings where they prevent flood damage. Sub-levees are used for under-seepage control. They also encircle areas behind the main levees which are subject to high uplift pressures during the high-water stages. Setback levees are generally built as backups to the existing levees. Spur levees project from the main levee; they serve to protect the main levees from the erosive effect of river currents (Fuhrman, 2000; Makhdoom, 2013).

Digital Elevation Model

This research uses the Digital Elevation Model (DEM) data to extract the possible locations of levee-like structures. The Digital Elevation Model data is also known as Digital Terrain Model (DTM) data. It is a statistical representation of the continuous surface of the ground by a lot of elevated points with X, Y, Z coordinates in an arbitrary coordinate field (Li, Zhu, & Gold, 2010). DEM maps are widely used to extract terrain parameters (Zandbergen, 2011) for geomorphologies such as levees (Fuhrman, 2000), hedgerows, bench terraces and ditches that greatly impact our living environment. Since levee-like structures have dimensions, only high spatial resolution data (i.e., with a pixel size of 16 square meters or less (Wulder, Hall, Coops, & Franklin, 2004)) is suitable for the classification process (Bailly, Lagacherie, Millier, Puech, & Kosuth, 2008). However, no matter how good the resolution is, the images are often heterogeneous and speckled due to the "noise" and artifacts caused by the sensor scanning. The main source of "noise" is caused by the angle of reflection, because the intensity values may be different from land to land as the angle of reflection varies (Song, Han, Yu, & Kim, 2002).

Imagery Noise

The imagery "noise" can be represented as the false positive data. In binary classification, the false positive is an error in data reporting. It improperly indicates the presence of a condition during a test. For example, the proposed classification method is supposed to only extract levees from the DEM, but the result contains roads and sand dunes. During the classification, the "noise" is commonly identified as a positive data, where it is actually the false positive data. In this study, the true positive data represents the preserved levee-like structures during the classification. The true negative data is the "noise" that was identified as "noise". The false negatives are the levee-like features that were classified as "noise". The false positives are the preserved "noise".

Purpose

There are existing flood management programs, such as Risk Mapping from FEMA, and Levee Safety Program that was established by USACE (U.S. Army Corps of Engineers). These programs assess and communicate risk to encourage better-informed decisions about the best flood risk reduction measures by individuals, businesses, levee sponsors, other responsible individuals and agencies. However, not all the levees are regulated by government agencies. For instance, the levees included in the Levee Safety Program account for only about 10% of the nation's levees (as estimated by National Committee on Levee Safety). These non-levee embankments are usually man-made agricultural embankments built by farmers and road and railroad banks (Olson & Morton, 2013). They are levee-like structures but are not certified or engineered to provide reliable flood protection. People may falsely believe these non-levee embankments can provide flood protection, and reside near them. Therefore, it is important to know where these features are located. This research aims at detecting and filtering all types of levee-like structures. Knowing the possible locations of levee-like structures will help to initiate better risk prevention plans.

LITERATURE REVIEW

For the past decades, Digital Elevation Models have been extensively used to extract terrain features. They are widely used in cartography, geographic information science, hydrologic modeling, and geomorphology where analysis can be done using only computers instead of digitizing, measuring topographic maps (Tarboton, Bras, & Rodriguez - Iturbe, 1991). This section highlights the past attempts of imagery extractions and imagery filter techniques over the past few decades.

In Peucker and Douglas' early research, a method to detect drainage network from DEM was created. The method marked the highest elevation point within each of the four adjacent points in the DEM, and all of the points which did not have marker were estimated as part of the drainage system (Peucker & Douglas, 1975). The theory behind this method is "a map of the concave-upward portions (low elevation points) of a DEM could be considered to be an approximation of the drainage network" (O'Callaghan & Mark, 1984). The algorithm did produce a reasonable result, but the extraction is strictly local data and the features are fragmentary. To avoid the same issue, O' Callaghan and Mark's method added abilities to filter imagery "noise", and extract more consistent drainage networks. The filter technique is an elevation smoothing process which is also known as low pass filter. Low pass filter is employed to remove high spatial frequency "noise" from an imagery by a moving window. The moving window affects one pixel at a time, changing its value by some function of local pixels that are covered by the window (Hong, 2016). According to the Matlab tutorial, choosing a right size of the kernel is very important because if it is too large, it may blur and remove small features of the image; if it is too small, it cannot eliminate "noise". O'Callaghan and Mark's used a 3 x 3 moving window for the smoothing process, but there is no specification on whether different moving window sizes were tested. Therefore, the result might be biased.

Filtering imagery noise has been a popular research topic over the past decades. Bayesian analysis is one of the most popular methods to minimize the probability of misclassification. The method is useful to estimate the distribution of the unknown parameters. The distributions can be applied to develop confidence intervals for the unknown parameters, and establish the significance of the estimated result (Dobigeon, Tourneret, & Chang, 2008). Mandy and his colleagues published a research about predicting landslides occurrence at the catchment of Payette River, Idaho. They found that both the Bayesian modeling and Chi-square analysis were useful on predicting mass movements. Nevertheless, Chi-square analysis only indicated variables that were related to the landslides occurrence, where Bayesian modeling was capable of indicating which variable has the most impact and how effective the significant variables were in predicting landslide location (Gritzner, Marcus, Aspinall, & Custer, 2001). In my study, elevation is the most relevant variable to the levee-like structures. My attempt focused on locating points that are significantly higher than the mean surrounding elevation through Bayesian statistics.

The second part of this study is eliminating “noise” from the extracted vector data (levee-like structures). Over the past years, there have been many studies about filtering morphological features from the LiDAR imagery, but surprisingly little has been published about filtering levee-like structures in a shapefile format.

On Baker and Cai’s early studies, they realized that Geographical Information Systems do not have many programs to calculate traditional measures of landscape feature (Baker & Cai, 1992), so they developed a program to analyze landscape structure by using a GIS software. One of the main methods of their research was to use the Minimum Bounding Circle (MBC) in the analysis of feature attributes, feature size, shape, dimension, and perimeter. The method compares the area of the feature to the area of the smallest circle that can circumscribe the feature (Baker & Cai, 1992). This method is particularly useful to identify features that are both linear and narrow (McGarigal, 2014). According to the Fragstats manual, a higher ratio (1 minus the quotient of the area of feature divide by the area of its smallest circle) generally indicates greater shape complexity from simple Euclidean geometry (McGarigal, 2014). For example, a sand dune can be as simple as a circle shape from the DEM-extracted data, and a levee-like structure can be more complex in shape. Since the outstanding characteristics of levee-like structures are long and linear, the MBC method can be applied to distinguishing levee-like structures. However, the limitation of this method is that, when patches are not completely extracted due to the inconsistent elevation level along the levee peak, this method cannot 100% preserve levee-like data because the removed false negative features do not satisfy the characteristics of levees. This kind of issue is especially common when the targets are natural levees because the shapes of natural levees are different due to water erosion.

Even though features with higher ratio are most likely levee-like structures, it is still hard to determine where the cut-point is. There are many methods to locate the optimal threshold, such as efficiency, misclassification-cost, odds ratio, kappa index, and the Youden Index. Further parameters like the decision costs and prevalence rates are not required by the Youden Index, a factor that makes it the easiest method for application (Fluss, Faraggi, & Reiser, 2005). The Youden Index (J) is formally defined as J = Maxc (Se (c) + Sp (c) - 1) (Fluss et al., 2005; Ruopp, Perkins, Whitcomb, & Schisterman, 2008; Schisterman, Perkins, Liu, & Bondell, 2005). It represents that if a cut-point achieves this maximum because the cut-off value optimizes the biomarker’s classification ability (Ruopp et al., 2008). Youden Index has usually used with ROC (Receiver Operating Characteristic) analysis. The Receiver Operating Curve is defined as the sensitivity versus 1-specificity plot over all the possible marker threshold values (Fluss et al., 2005). The sensitivity and specificity are the probability of truly identifying diseased and non-diseased individuals respectively at a certain threshold value (c). The precision of threshold value can be determined by the probability of a sensitivity and the probability of a specificity (Fluss et al., 2005). The ROC graphs and Youden Index are commonly used in medical field, but both methods have been increasingly used in other areas, such as machine learning and data mining research (Fawcett, 2006).

In the medical field, the Youden Index and ROC curve are usually used to discriminate healthy and diseased individuals (Fluss et al., 2005; Ruopp et al., 2008; Schisterman et al., 2005). Similarly, these two variables can be replaced with levee-like structures and “noise” for my research. For instance, an extracted feature is assessed as a levee-like structure if the MBC ratio is greater than a given threshold value, otherwise, the feature is a “noise”. The optimal threshold value can be located when equal weight is given to sensitivity and specificity (Ruopp et al., 2008). One key factor determines the accuracy and efficiency of the optimal threshold, and that is the sample data. Marcus suggested that it is important to have a good amount of sample data because more sample data generally increases precision when predicting unknown parameters (Ruopp et al., 2008). For example, an optimal threshold value located through 100 surveyed features is likely more significant than the one located through 10 surveyed features. Insufficient sample data frequently introduces bias and inaccuracy into the estimates of parameters (Ruopp et al., 2008).

By the inspiration I derive and in a bid to advance on all these studies, I came up with my own method on the non-embankment detection project. The proposed method attempted to use Empirical Bayesian theorem and the low pass filter techniques to extract elevated linear features from Digital Elevation Model (DEM), and combine the Minimum Bounding Circle (MBC) method and the Youden Index to locate the optimal threshold value that can be used to determine “noise” and levee-like features. The expectations of this study are:

1. Extract elevation points that are most relevant to levee-like structures, and convert the result to shapefile format.
2. Compare the survey data against the extracted features, and find evidence to prove if the MBC ratio is useful to differentiate levee-like structures and “noise”.
3. Use the Youden Index and ROC curve to locate the optimal threshold value.
4. Filter the extracted data with the optimal threshold value, examine the accuracy and efficiency of the result.
5. Test if the optimal threshold value can be used to filter other counties.

DATA SOURCE

Three counties DEM data were used in this study, Morgan County, Pulaski County, and Lawrence County (Figure 2). The DEMs (IMG format) were downloaded from the Indiana Spatial Data Portal (http://gis.iu.edu). The Mosaics were created by Chris Morse, Natural Resource Conservation Service. The horizontal resolution is 5 feet. The horizontal and vertical units are in US Survey Feet. The projection used is the predominant Indiana State Plane zone for a county on NAD 83 datum and NAVD88 (GEOID99) vertical datum.

The sample data (levee-like structures) was provided by Polis center in ESRI shapefile format. The data is collected with GPS unit from study areas of Morgan County, Indiana, and Pulaski County, Indiana.

Pulaski County is located at the North West of Indiana. According to the 2010 census, the county has a total area of 434.53 square miles; 433.65 square miles are land and 0.88 square miles is water. The average elevation is 705 feet. The land is nearly flat, except for some slightly sloping terrain along the Tippecanoe River and Tributaries to the east (Figure 3a).

Lawrence County is located at the South of Indiana. The county has a total area of 451.93 square miles, of which 449.17 square miles is land and 2.76 square miles is water. The average elevation is about 502 feet. Although lying at the lower altitude, it features more hills than the northern areas in the state (Figure 3b).

Morgan County is located in the Central Indiana. It contains a total area of 409.43 square miles, of which 403.97 square miles is land and 5.46 square miles is water. The elevation is 624 feet above the sea level. It is within the Till Plains region, which lies to the south of the Great Lakes Plains. The landscape of this region is characterized by low hills and valleys (Figure 3c).

There are 5 major rivers running through Indiana State: Ohio River, Kankakee River, White River, and Tippecanoe River. The White River runs through Morgan and Lawrence counties, and the Tippecanoe River runs through Pulaski County.

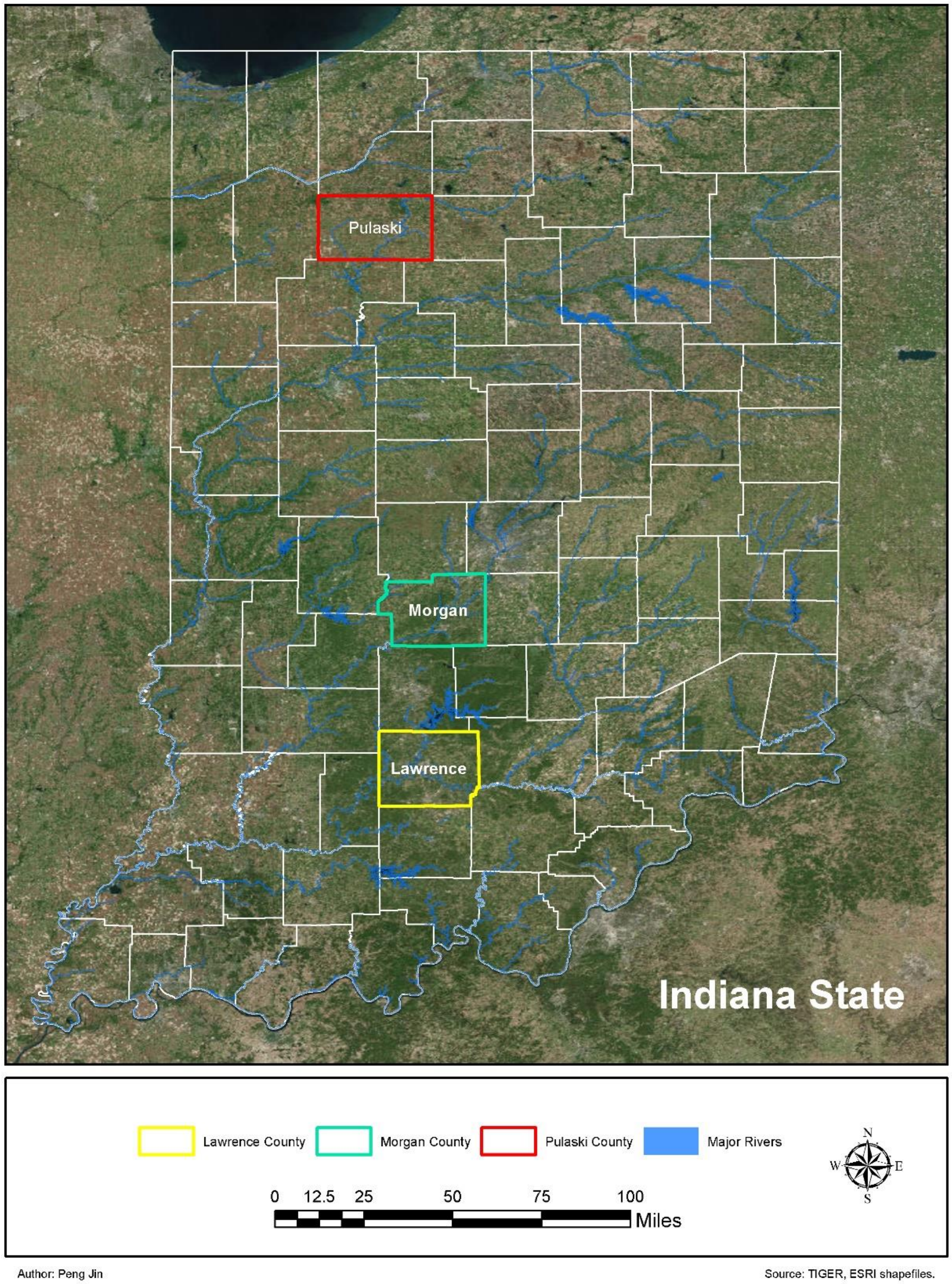


Figure 2: *The geographical locations of Pulaski County (North West), Morgan County (Central), and Lawrence County (South) in Indiana State. Scale 1:2,500,000*

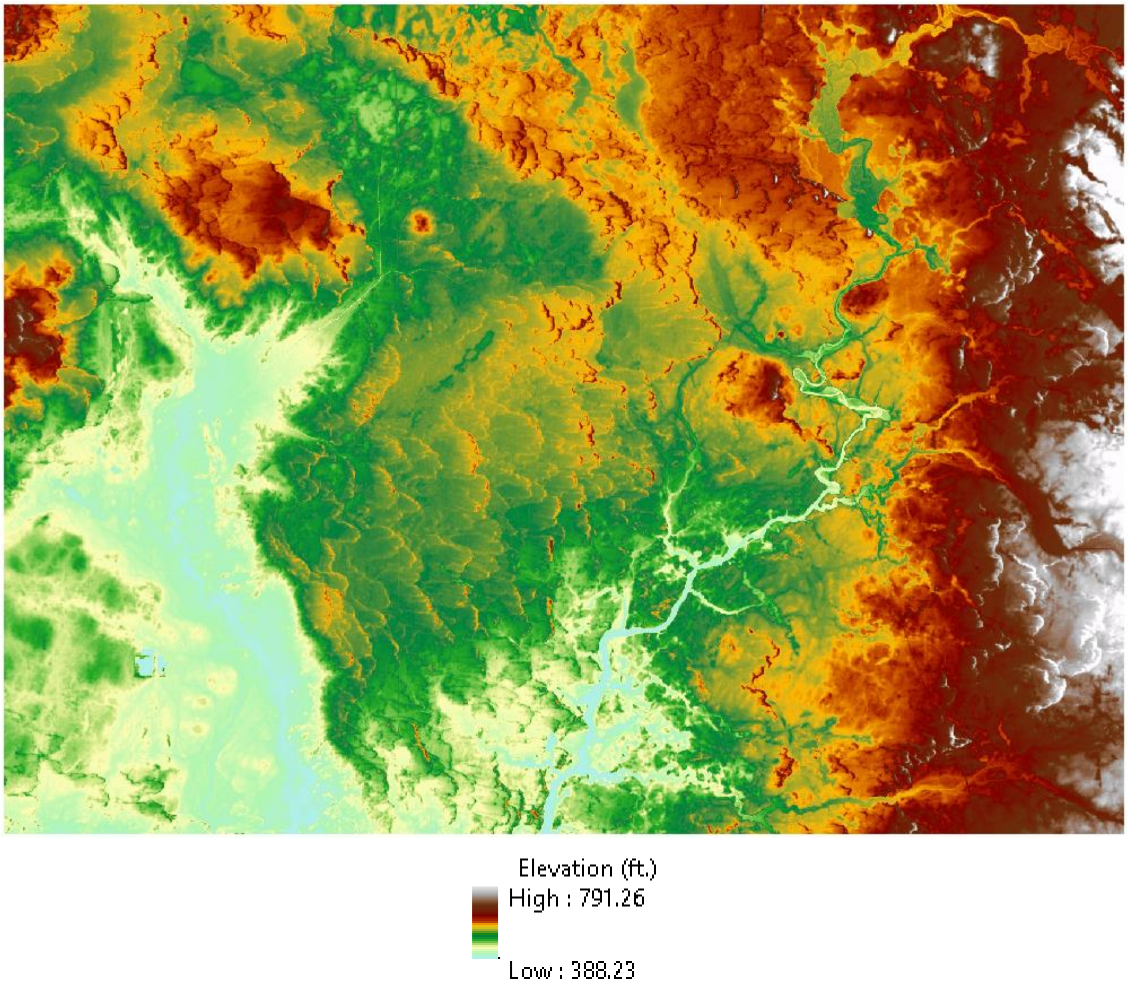


Figure 3a: *Pulaski County – The image was created through ArcMap, and symbolized by the elevation value. The Tippecanoe River flows from northeast to the south. The west part of the map appears as blue color due to its lower elevation level (no major river systems). As the map shows, there are more flat areas across the whole county. The elevation difference from low to high is 403.03 feet.*

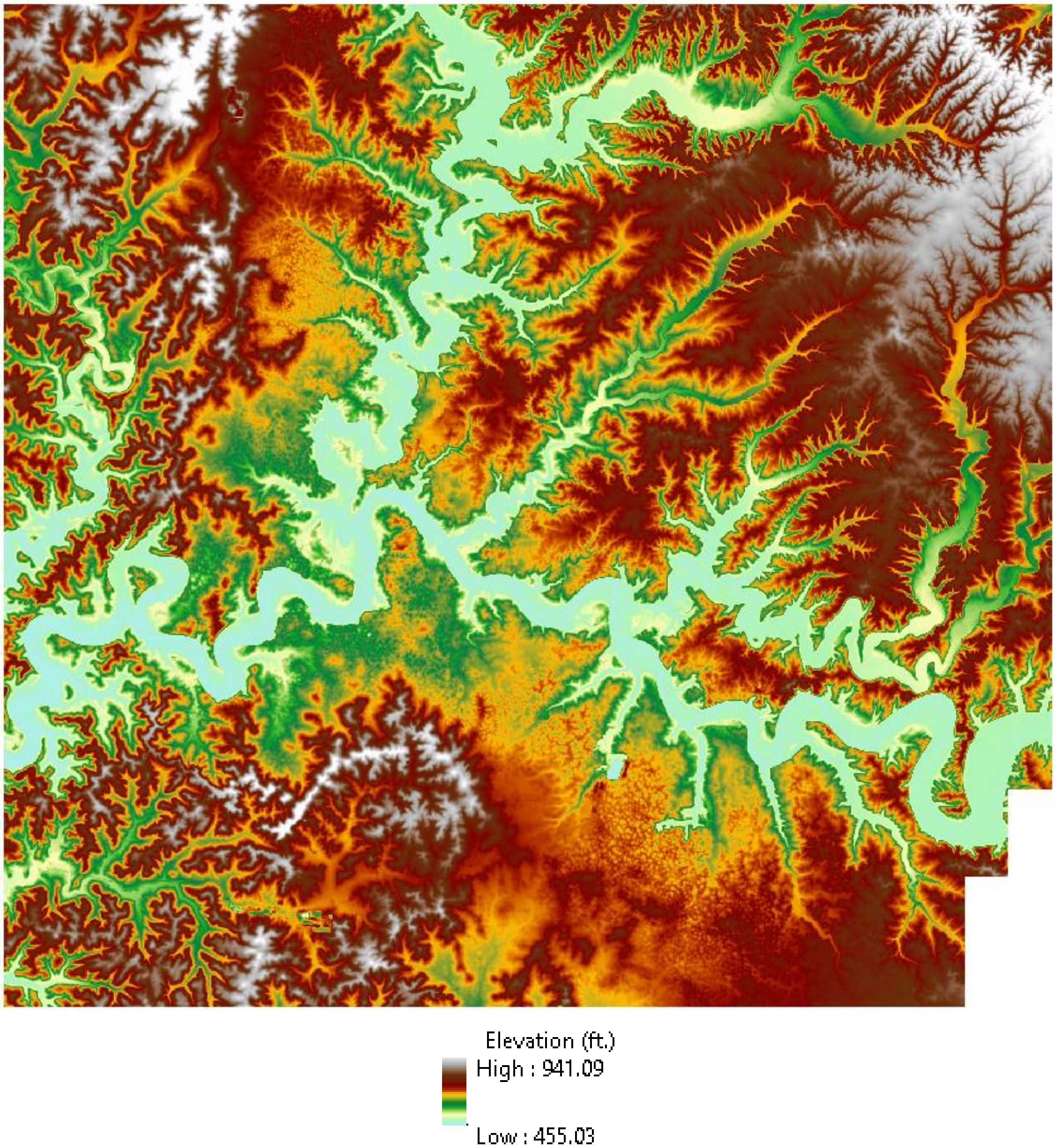


Figure 3b: *Lawrence County – the overall elevation of Lawrence County is higher than Pulaski County. The light blue area represents White River system. There are more high elevation areas in this county. The elevation difference from low to high is 486.06 feet.*

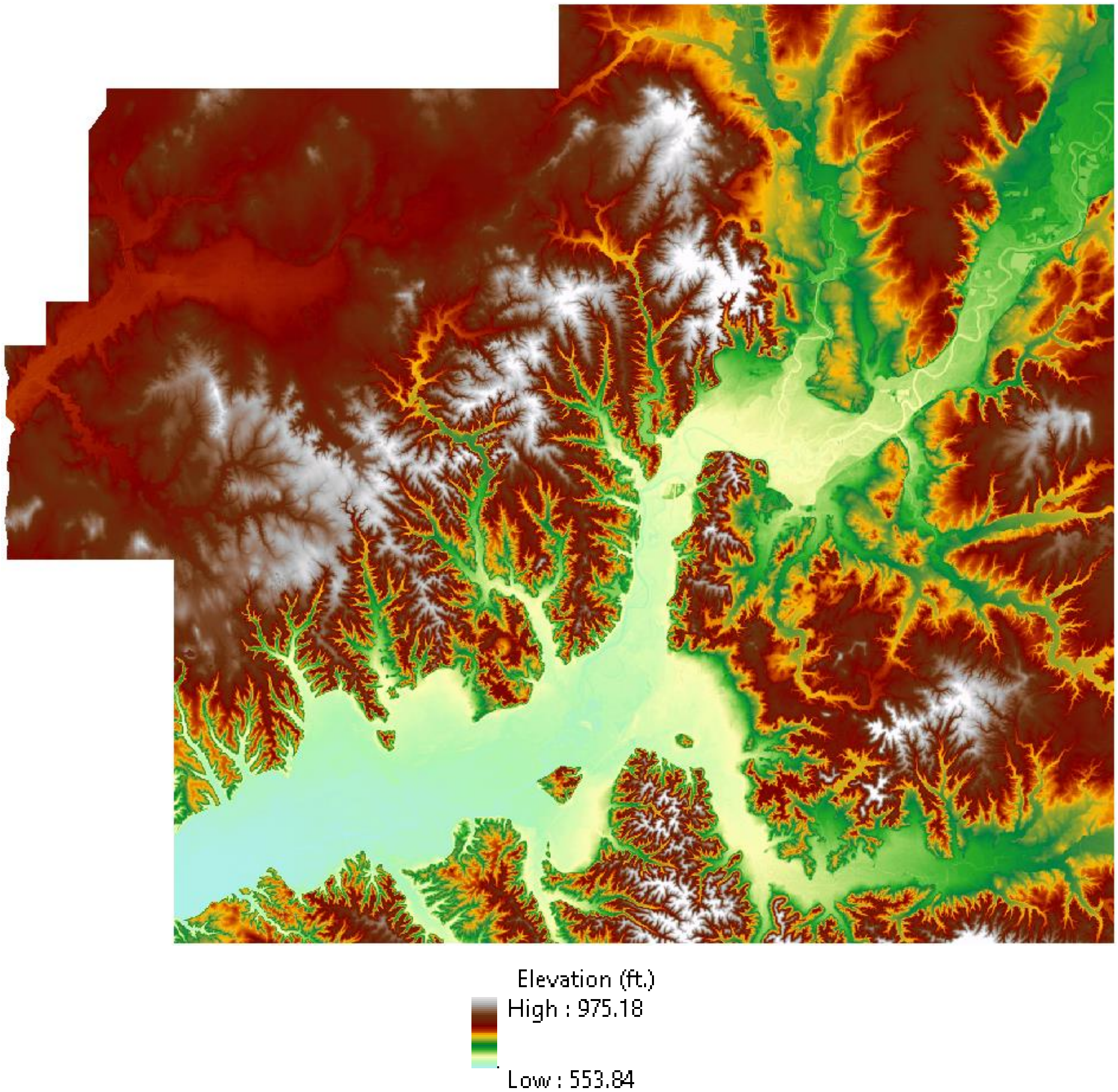


Figure 3c: *Morgan County – Morgan County's landscape appears to have both the other two counties' characteristics. The distribution of flat and high areas is about the same. The linear features within the green symbolized areas are river channels. The elevation difference from low to high is 421.34 feet.*

METHODLOGY

Workflow

Figure 4 shows the general workflow of this research.

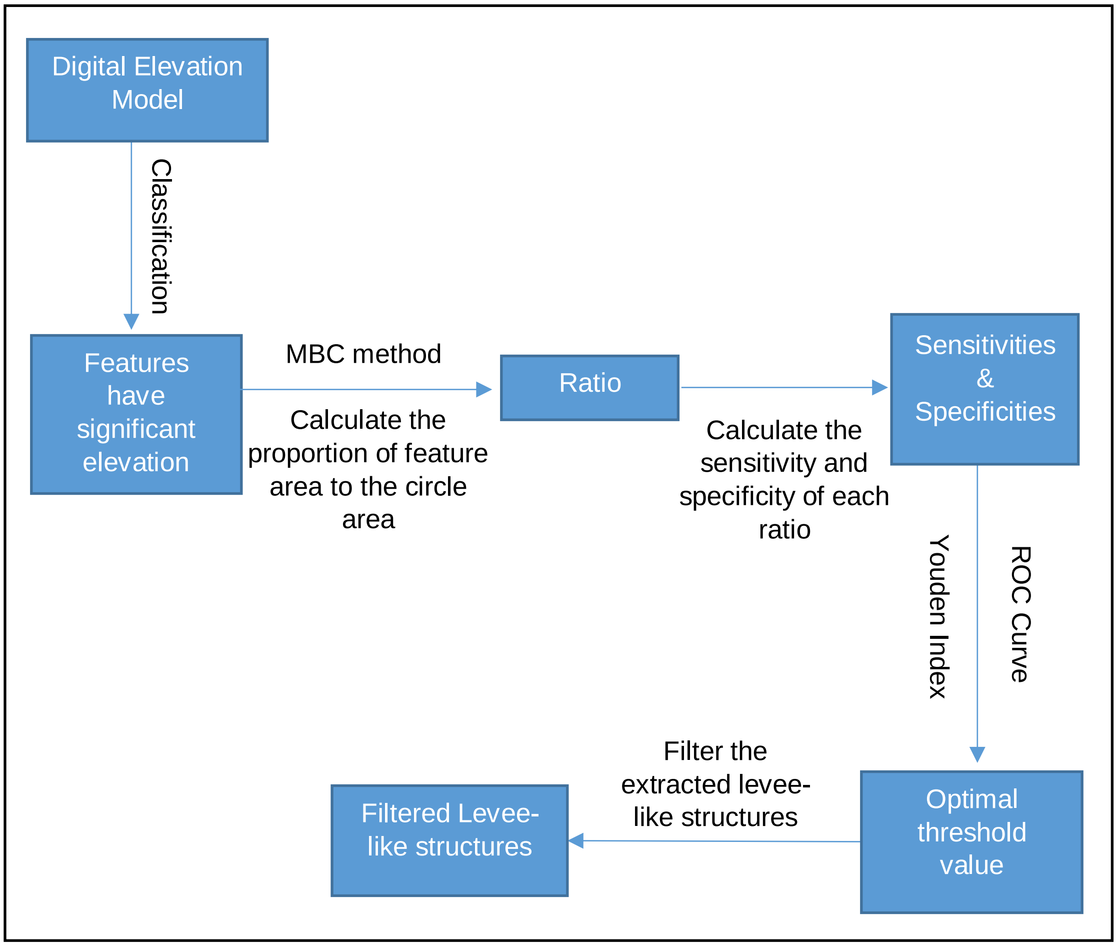


Figure 4: *Workflow of morphological filters on DEM.*

Imagery Classification

The first proposed method was a Bayesian statistic based program that was written in the R programming language. This section classifies levee-like structures from the digital elevation model (DEM). Below paragraph is the line by line interpretation of the script.

1. The start of the program loads a DEM image to the system and assigns the image file to a variable as the observed elevation data. 2. It then runs an 11X11 moving window on the original raster to calculate the expected elevation around

the observed elevation. After comparing the results that were generated by different moving window sizes, the 11 x 11 moving window produced the best result for this study. 3. The program then calculates the standardized elevation ratio by using the observed elevation divide by the expected elevation. 4. To analyze the spatial relationship between each elevation points and its 8 neighbors, the program runs two focal, moving window operations to calculate the variance across the observed raster and the variance in the expected raster. This step performs a nearest neighbor analysis to determine variance by using 3 x3 moving windows, where the center pixel that gets the value of the variance of all the pixels surrounding it. 5. Two variables are calculated for the calculation of Empirical Bayes, global mean of standard elevation ratio, and variance ratio between observed variance and expected variance. 7. The program uses Empirical Bayes theorem classifies features (levee-like structures) that are statistically higher than other morphological features by using (shrinkage factor x standard elevation ratio) + ((1-shinkage factor) x global mean). 8. In order to generate the classification matrix to further identify levee-like structures, the program exports the classified data to an image file (.img file extension), and create a new classification matrix based on the characteristics of levees. 9. The script then uses the new matrix to reclassify the processed raster, and output the filtered levee-like structures as an image file. This step exports the most significant portion of data that reduces any "noise" generated by the method. 10. The last step was done by using the *RasterToPolygon* tool from ArcMap. It converts the raster file to an ESRI shapefile for the next filter process.

Minimum Bounding Circle

The second part of this study used MBC (Minimum Bounding Circle) method to calculate the patch ratio and later used the ratio to differentiate levee-like structures and noise. The areas of morphological features were calculated through the "Calculate Geometry" tool from ArcMap 10.2. The unit is in U.S. square feet. The diameters of circles were automatically generated when running the "Minimum Bounding Geometry" tool. The circle area was calculated by using the circle area formula (πr^2). The ArcMap ratio was calculated by using the quotient of the area of morphological feature divide by the area of the smallest circle that can circumscribe the feature (Figure 5).

Related Circumscribing Circle	
Circle = 1 – [$\frac{a}{b}$]	a = area (m^2) of the patch b = area (m^2) of the smallest circumscribing circle around the patch

Figure 5: *Circle ratio equals to 1 minus patch area divided by the area of the smallest circumscribing circle. 0 <= CIRCLE < 1. Note, this index is not influenced by the patch size (Baker & Cai, 1992; McGarigal, 2014).*

Since the compactness of the shape approaches the boundaries of the circle (Baker & Cai, 1992), the circle area should be greater than or equal to the area of the feature. Therefore, the ratio between the feature and the circle is within 0 to 1. The ratio between a levee-like structure and its minimum bounding circle is usually a very small number because levees are usually long linear structures, so the areas of features are generally smaller than their minimum bounding circles. Whereas a non-levee structure is less complex, and its area is closer to the circle area, so the ratio is usually a greater number (Figure 6). Thus, the smaller the ratio, the greater the possibility that the feature is a levee. In figure 5, the formula uses 1 minus the ratio to reverse the result (McGarigal, 2014). It indicates that the less complex the patch is, the lesser the chance it is a levee-like structure. The range of reversed ratio is still between 0 to1. It is just a different way to represent the same set of data. Figure 7a is a study area in Morgan County, Indiana. The red features are surveyed levees. The green segments are DEM-extracted morphological features. The circles are minimum bounding circles to the patches. Each patch has a ratio value that was calculated with MBC method. Figure 7b and 7c illustrate how MBC method discriminates levee-like structures and “noise”.

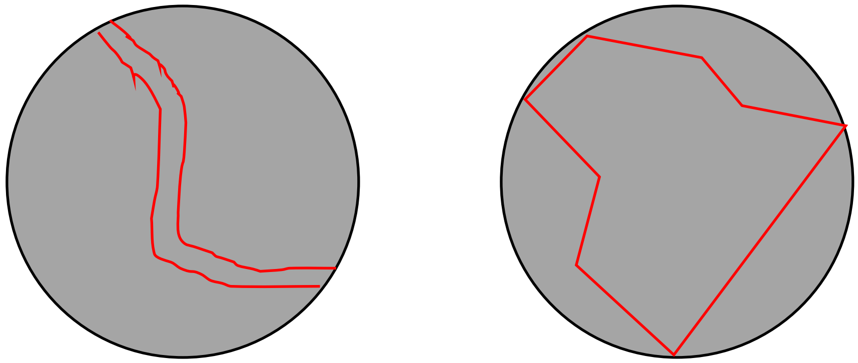


Figure 6: *Levee-like structure (left), Non-levee feature (right)*

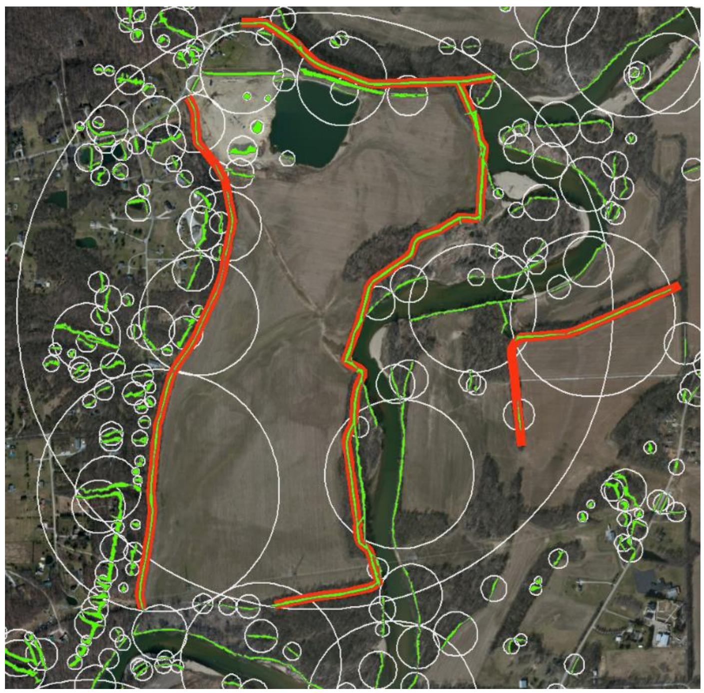


Figure 7a: *This is a study area in Morgan County, Indiana. Surveyed levee-like structures are in red color. Green segments are extracted morphological features from DEM. Each morphological feature has a minimum bounding circle around it. The survey data was provided by the POLIS Center, IUPUI. Scale 1:12,000.*

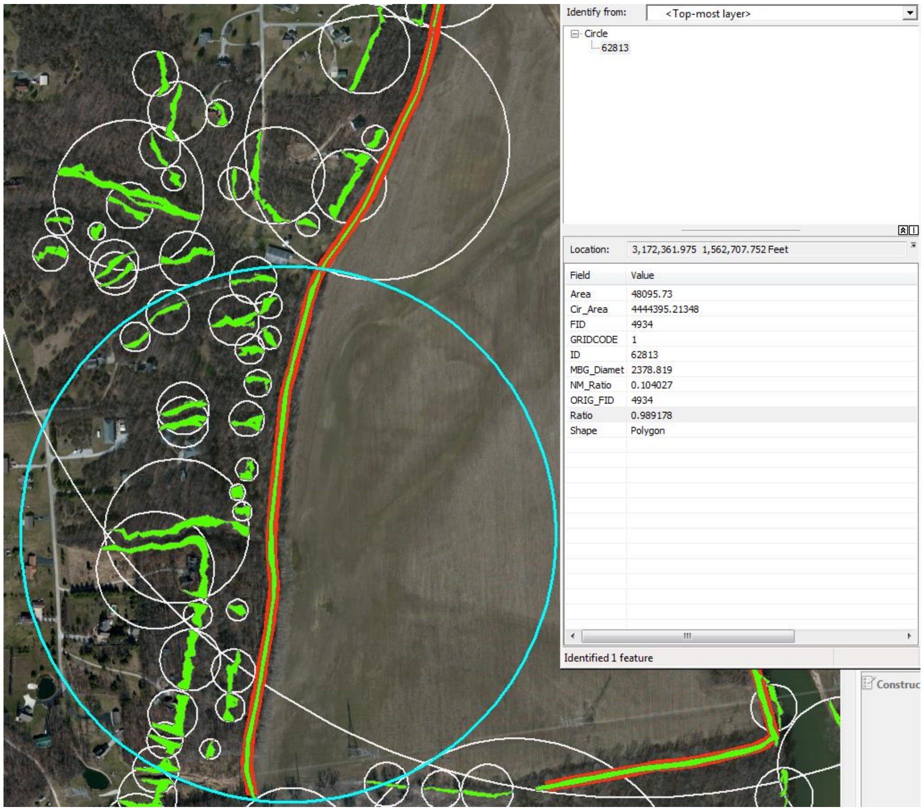


Figure 7b: *The table shows the area of selected levee feature is much smaller than its MBC area, so their proportion is small (48095.73 / 4444395.21348 = 0.010822). Scale 1:6,000.*

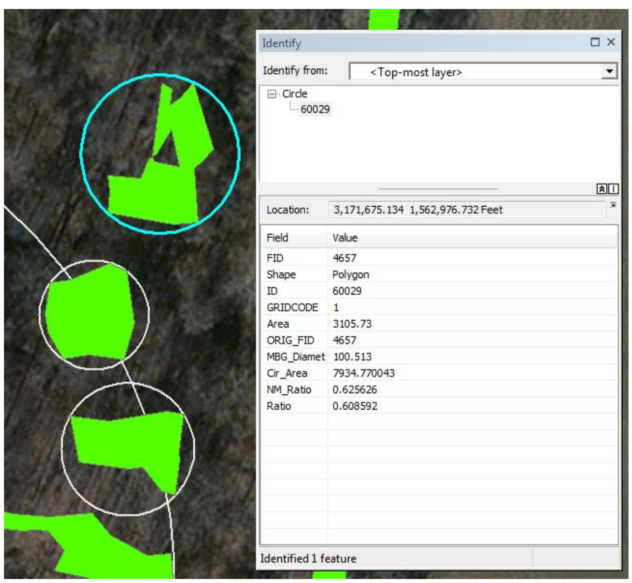


Figure 7c: *This screenshot illustrates the false positive feature and its MBC. The circled patch shape is more regular than a levee-like structure, and the shape area is close to the MBC area, so their ratio is greater number compare to a levee-like structure (3105.73/7934.770043 = 0.391408). Scale 1:500.*

The data shows levee-like structures tend to have smaller ratios, and "noise" features likely have greater ratios. Therefore, MBC ratio is a valid method to determine if the extracted morphological features are levee-like structures or "noise". The question is, how do I locate the cutoff point? The next method explained how efficient to Youden Index and Receiver Operating Characteristic curve to find the optimal threshold value.

Youden Index & Receiver Operating Characteristic

The Youden Index is a commonly used summary measure of the Receiver Operating Characteristic (ROC) curve. This index can be defined as *J = max (sensitivity c + specificity c – 1)*, the range is between 0 and 1. J provides a criterion for choosing the "optimal" threshold value (c*), the threshold value for which *(sensitivity + specificity – 1)* is maximized (Fluss et al., 2005). A value of 1 represents no false positives or false negatives in the result, i.e., the result has no "noise". In the ROC curve, all cut-off points are represented by plotting the true positive rate (Specificity) and false positive rate (1-Specificity). A sensitivity/specificity pair that corresponds to a given specific decision threshold is represented by every point on the ROC curve. Correctly predicted true positive results are expressed as a percentage which is the sensitivity (Powers, 2011). It is useful for ruling out the "noise" data. It refers to the method's ability to correctly detect levee-like structures. It can be expressed as:

$$Sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

$$= \frac{number\ of\ true\ positives}{total\ number\ of\ levee{-}like\ structures\ in\ the\ area}$$

$$= probability\ of\ the\ identified\ features\ is\ levees$$

A high sensitivity result is dependable when the data is negative since it is not likely misidentifying those non-levee data. If the output has 100% sensitivity, all extracted data will be levee-like structures.

Specificity is the percentage of true negative results that are correctly predicted (Powers, 2011). It relates to the method's efficiency to correctly detect true negative features. It can be expressed as:

$$Specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}$$

$$= \frac{number\ of\ true\ positive}{total\ number\ of\ non{-}levee\ like\ structures\ in\ the\ area}$$

$$= probability\ of\ the\ identified\ non\text{-}levee\ data\ is\ not\ levee$$

It is useful for ruling in non-levee like structures when a positive data in the result with high specificity because the method rarely identifies positive data in non-levee data. A result with 100% specificity indicates the non-levee structures are accurately excluded.

I wrote a simple program to automatically calculate and compare all the sample data. The program was written in Perl (Practical Extraction and Reporting Language). All surveyed features (levee-like structures) were considered as true positives, and everything else in that area was considered as a false positive feature. The data was saved in a text file which contained two columns, ratio, and marker. The ratios were sorted in ascending order. The maker field indicates the feature is a true negative (marked as B) or a true positive (marked as A). The program takes each ratio as a cut-point regardless it is a true positive or true negative. Every A smaller than the threshold value of the time will be considered as a false negative data because the program assumes every data smaller than that threshold is a “noise”. Every B smaller than the cut-point of the time is a true negative data, all As that greater than the threshold value of the time are true positives, and every B greater than the threshold value of the time is a false positive data. The program calculates the sensitivity and specificity at the end by using the count of As and Bs in each group. Here is an example to explain how the program works, if the program finds 3 As and 7 Bs smaller than the threshold value of the time, and 6 As and 2 Bs greater than that threshold value, there are 3 false negatives, 7 true negatives, 6 true positives, and 2 false positives. The program calculates the sensitivities and specificities of all the ratio at the end. Lastly, the program generates an output in text file format that contains sorted ratio, marker, sensitivities, and specificities. By importing the output into an EXCEL spreadsheet, it allows me to use Youden Index function (sensitivity + specificity-1) to calculate all the threshold values, and locate the optimal threshold value which is the greatest number within the threshold column.

A full script is available upon request.

Figure 8 is a demonstration of the output in a ROC curve. Theoretically, the perfect cut-point should be located at the upper left corner or coordinate (0,1) of the curve. It indicates 100% sensitivity and 100% specificity. However, coordinate (0,1) is only an ideal result. It is impossible to have a 100% accurate cut-point for this study. The best result would be the closest coordinate at the upper left corner.

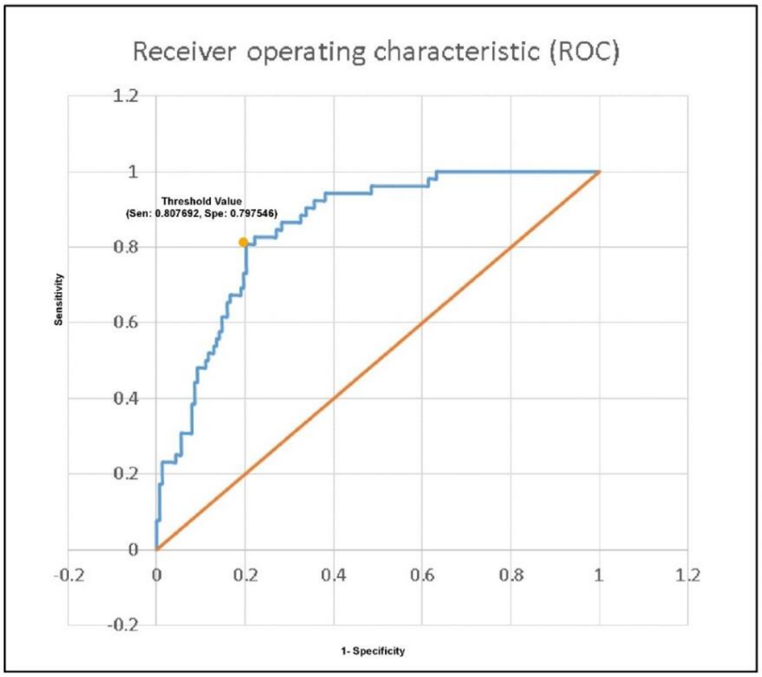
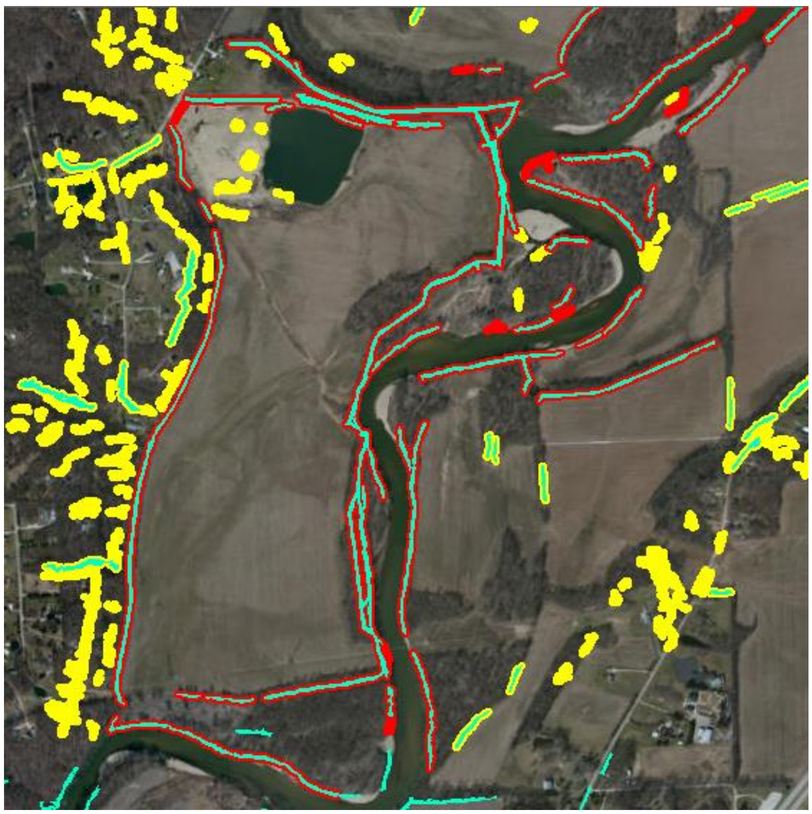


Figure 8: *The X axis is 1- specificity to display the data from small to large, so the data needs to be reversed. Y axis represents sensitivity.*

RESULT

For Morgan County, the corresponding ratio value to the optimal threshold is 0.904249026. Figure 9 shows the true positive features (segments in cyan color) are mostly preserved in the study area, but some "noise" features still exist. These structures are mostly road features, and some are levee-like structures that did not get surveyed. According to the U.S. Army Corps of Engineers' *Design and Construction of Levees*, some roads are built on top of levees, such as access road on levee (Fuhrman, 2000). Therefore, roads can be potential levees. Figure 10 shows the filtered result with symbolized DEM as the base map.



ROC Filtered Feature *True Positive Data* *True Negative Data*

Figure 9: *This is a surveyed area in Morgan County. Red features are surveyed levees (data acquired from POLIS center), yellow features are true negative data, and cyan features are preserved levees after the filter.*

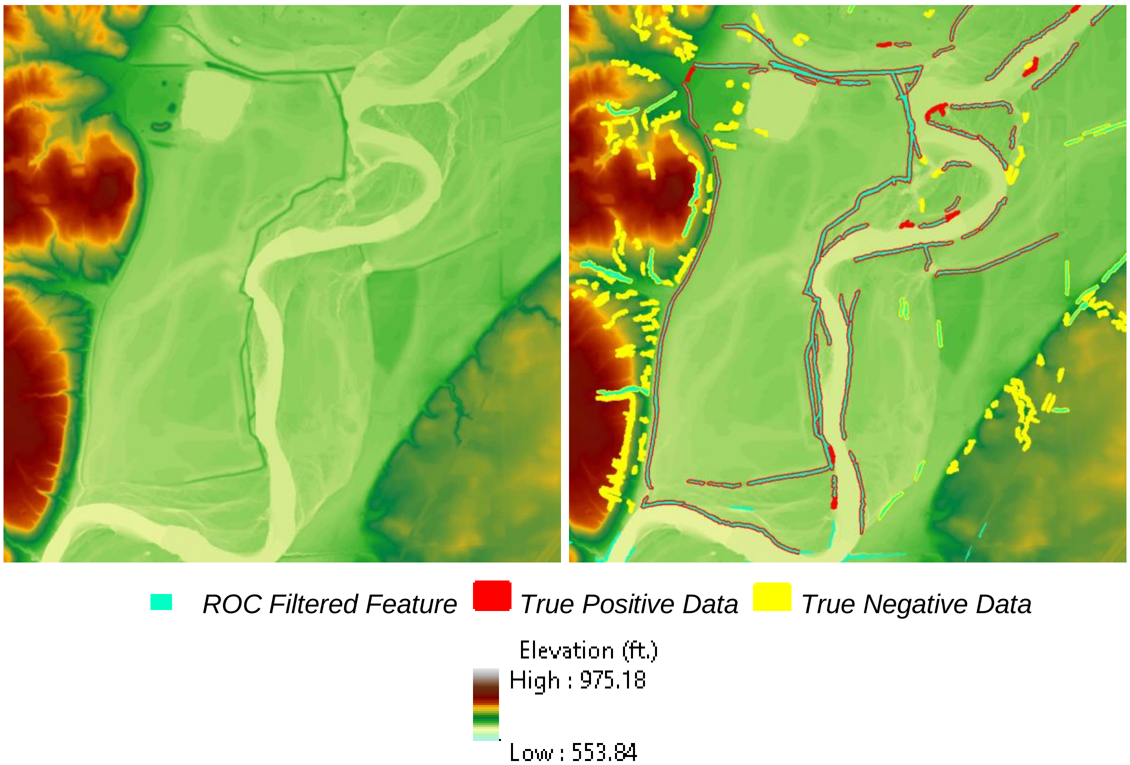


Figure 10: *The background map is a symbolized DEM imagery. The left screenshot shows a plain map with no overlaid features. The right screenshot overlaid with filtered levee-like structures (cyan color), survey data (red color), and "noise" (yellow color).*

Figure 11 shows the overall comparison between the raw data and filtered data in Morgan County. The raw data has 22,427 fragments, and the filtered data lower the number down to 5,737. It's about 74.4% of all the data. Undoubtedly, there are some false negatives (levee data) within the 74.4% of deleted data. The method is designed to maximally preserve levee-like structures that have obvious levee characteristics. The main cause of false deletion is the inconsistent extraction during the imagery classification process because some levee-like structures do not have steady height, so the data cannot be extracted as linear shapes.

The optimal threshold ratio shows 80.8% sensitivity and 79.8% specificity. Both rates are pretty high. It proved the method had 80.8% accuracy rate to preserve true positive data, and 79.8% accuracy rate to remove "noise". Figure 12a, 12b, and 12c show the comparisons between survey data and filtered results. Some legitimate levee-like features were detected by the extraction method, but they were not in the survey data. The chances are the survey progress wasn't complete at that time or those areas were not accessible.

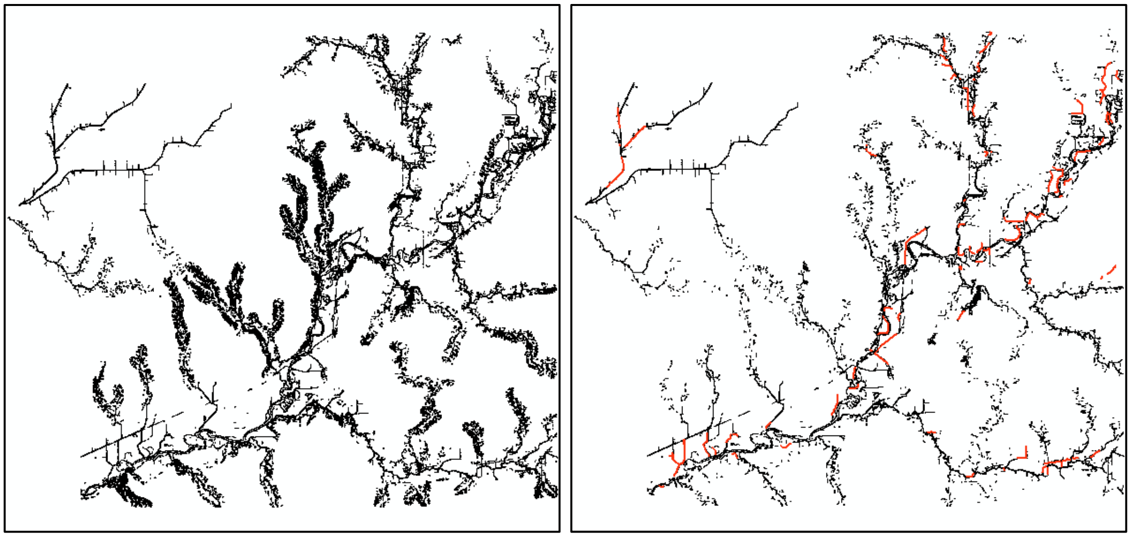


Figure 11: *Morgan County, before the filter (left) and after the filter (right) at scale 1:300,000. Red features are surveyed levees; black features are extracted levee-like structures.*

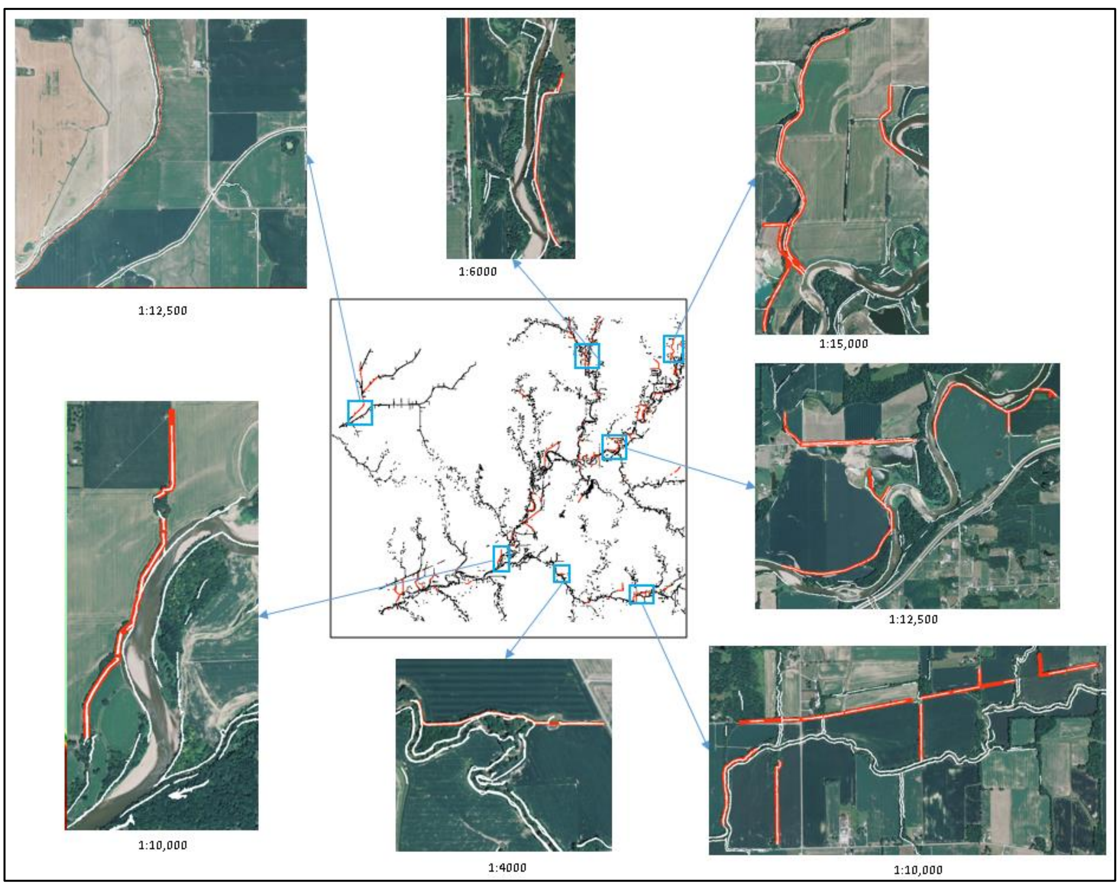


Figure 12a: *The filtered data overlaid on top of survey data with the aerial base map. Red features are surveyed levees, and white features are filtered levee-like structures.*

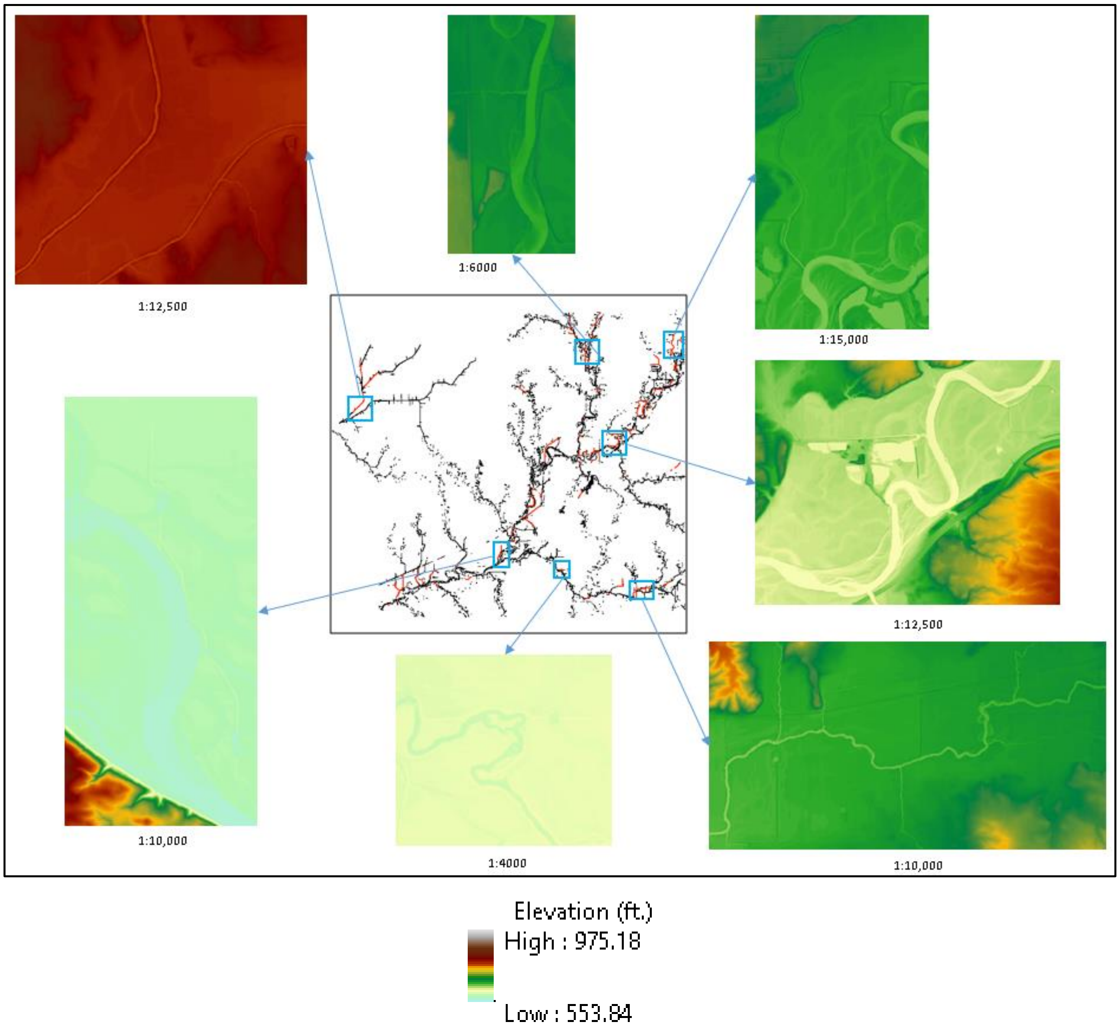
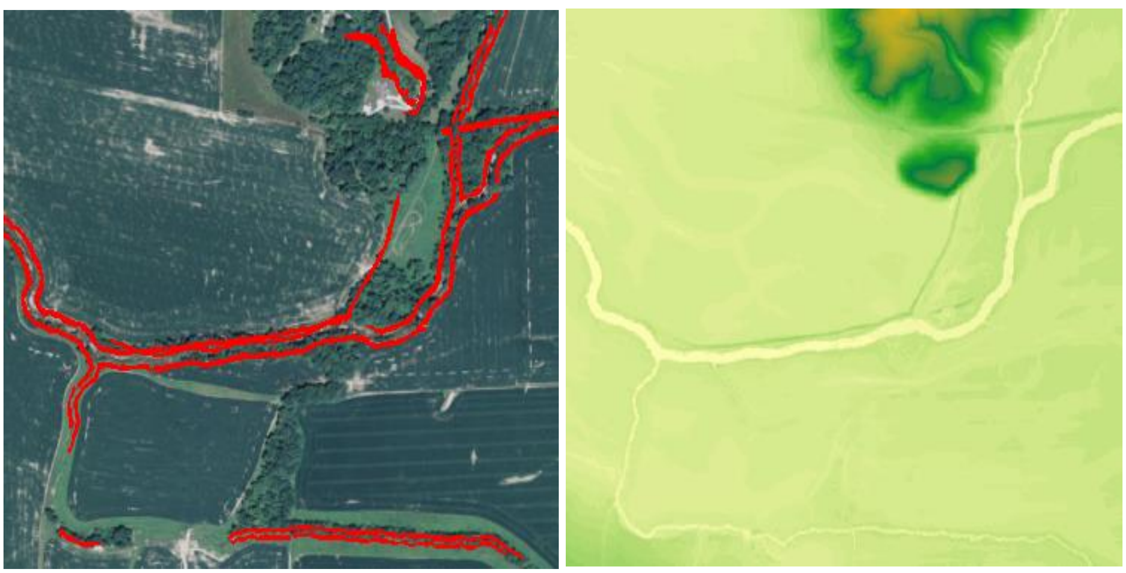
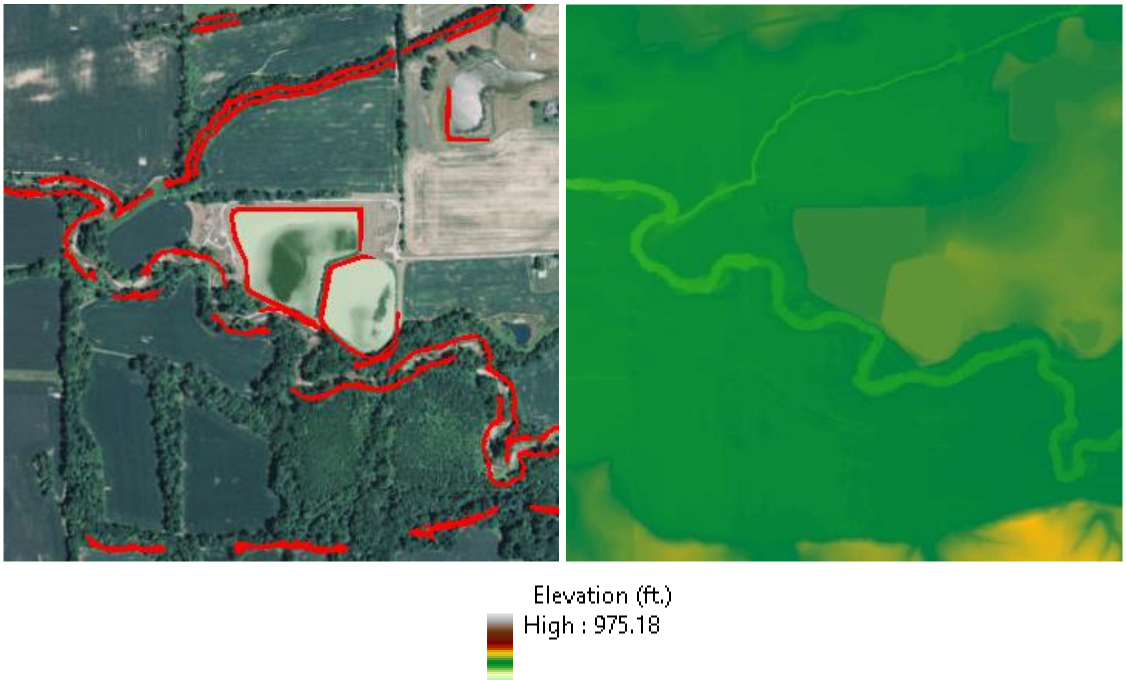


Figure 12b: *These screenshots are DEM representations of the same areas in figure 12a. The river boundaries can be clearly seen from the imagery. This figure provides a general review of the accuracy and efficiency of the filtered result.*

Scale 1: 20,000



Scale 1: 10,000



Scale 1: 10,000

Figure 12c: *Above are filtered results from randomly selected areas in Morgan County. Aerial imagery (left),and symbolized DEM (right).*

Comparing the filtered results to the symbolized DEM, most of the filtered features look very reasonable. The majority of the long linear features was preserved along the river channels. In order to test if the optimal threshold ratio of Morgan County is also capable of filtering other counties, I applied Morgan County's optimal threshold value to Pulaski County and Lawrence County's raw data.

For Pulaski County, the filtered result shows 72.62% sensitivity and 97.82% specificity with optimal threshold ratio of Morgan County, which indicates the threshold value provided 72.62% accuracy rate to preserve levee-like data, and 97.82% accuracy rate to remove "noise". The sensitivity of Pulaski County (72.62%) is slightly lower than Morgan County (80.8%) because there are more true positive data that has small ratios in the test area of Pulaski County. It is harder to extract terrains when the profile of the study area is flat (O'Callaghan & Mark, 1984), so the result is likely to contain inconsistent patches during the imagery classification process in Pulaski County. On the other hand, the specificity of Pulaski County's test is much higher than Morgan County's specificity rate. Overall, the cutoff point of Morgan County is a viable choice to filter Pulaski County's data.

Figure 13 shows the overall comparison between the raw data and filtered data. Figure 14 are filtered results from randomly selected areas.

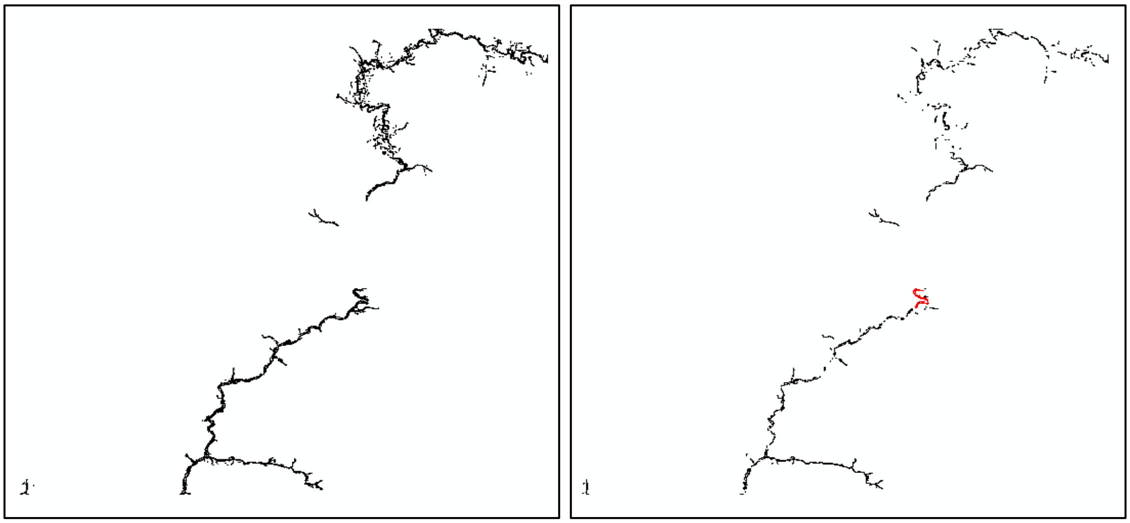
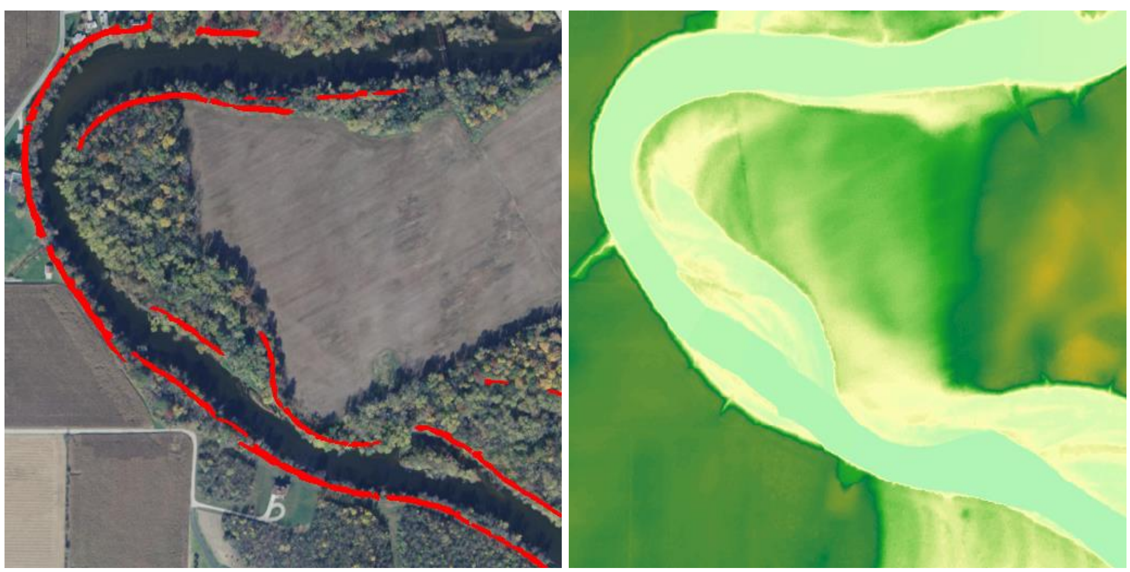


Figure 13: *Pulaski County, data before the filter (left) and after the filter (right) at scale 1:250,000.The red features are surveyed levees. The black features are extracted morphological structures.*



Scale 1: 6,000

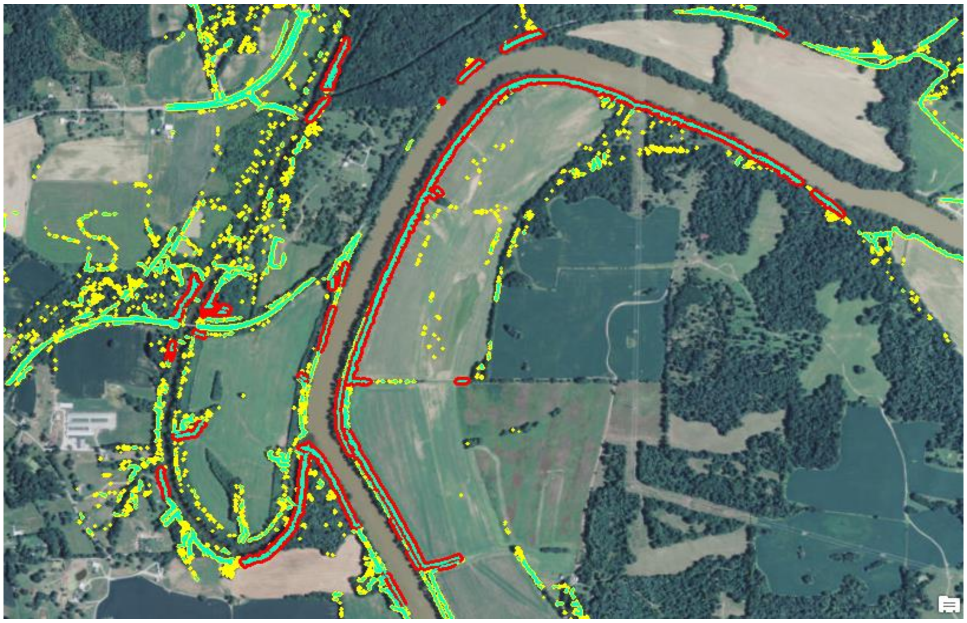
Scale 1:6,000

Elevation (ft.)
High : 791.26
Low : 388.23

Figure 14*: Above are filtered results from randomly selected areas in Pulaski County. Aerial imagery (left),and symbolized DEM (right).*

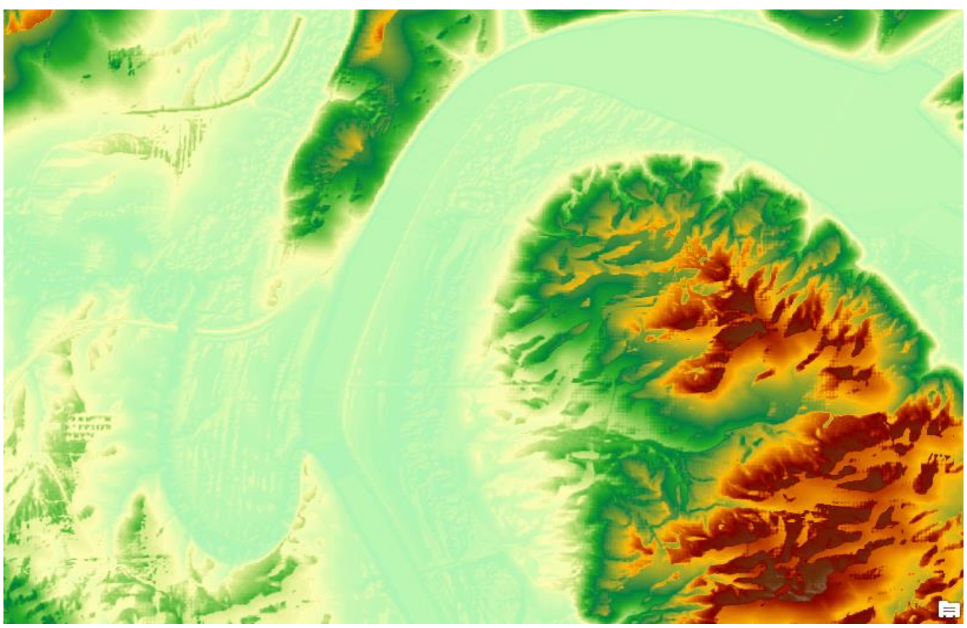
Unfortunately, the survey data in Lawrence County was not available. Alternatively, I found an area where I could manually identify levee-like structures based on my own knowledge. The sensitivity and specificity rates are around 71.58% and 99.28% when used the cutoff value of Morgan County to filter the extracted raw data. The result is about the same as that of Pulaski County (72.62% sensitivity and 97.82% specificity). 71.62% of sensitivity is also lower than Morgan County's sensitivity rate. The cause is also due to a large amount of inconsistent and irregular patches that exist in the data. However, if 71.58% of the whole county's levee-like structures could be preserved, the general outlines of all the levee-like features were already clear to see. Furthermore, the false negative data mostly contains small fragments, which are more in number, but less in size, so it does not have a big impact on the overall result. Figure 15a, 15b, 15c, and 15d show the filtered results from different aspects.

The results from both Pulaski and Lawrence County proved the optimal threshold ratio of Morgan County was efficient to filter other counties' data. Through the whole research, the methods produced efficient results for the selected counties, but there were also potential issues that need to be further discussed and studied.



ROC Filtered Feature Ture Positive Data True Negative Data

Figure 15a: *A test area in Lawrence County at scale 1:15,000. Red features are self-identified levees, yellow features are true negative data, and cyan features are filtered results.*



Elevation (ft.)
High : 941.09
Low : 455.03

Figure 15b: *A test area in Lawrence County – DEM representation.*

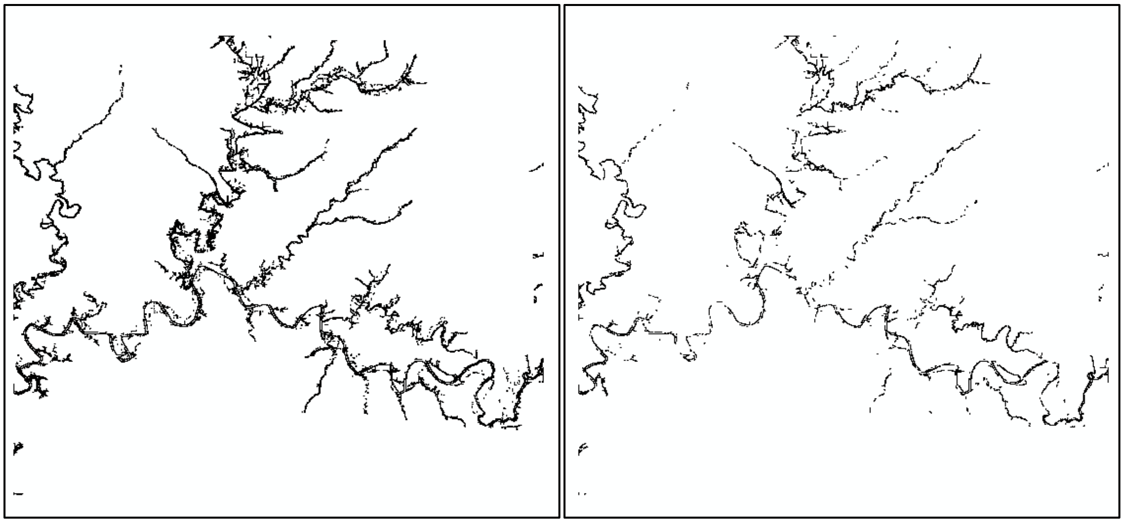
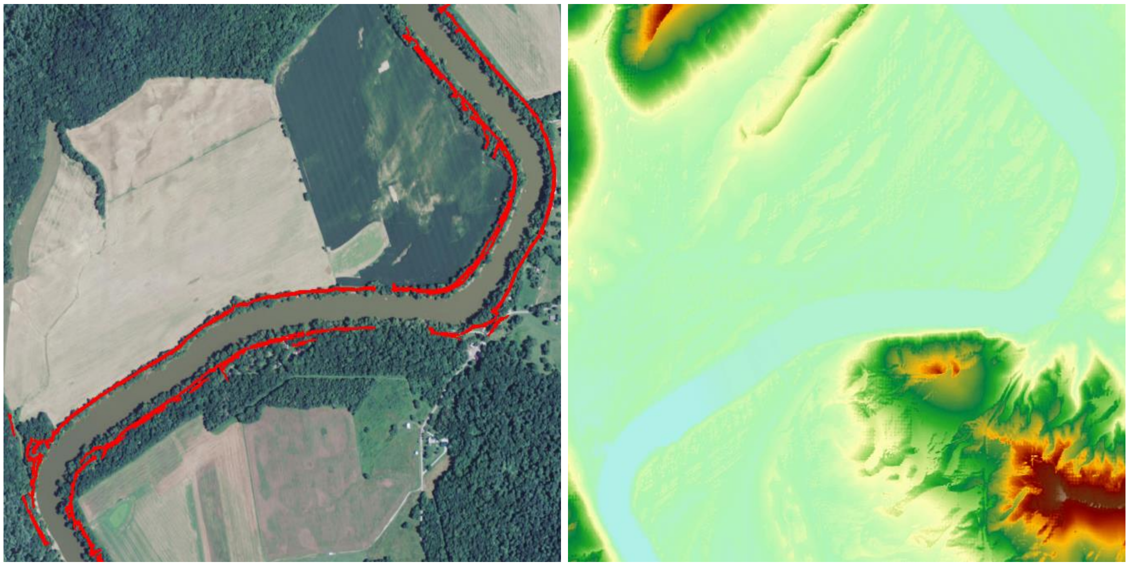
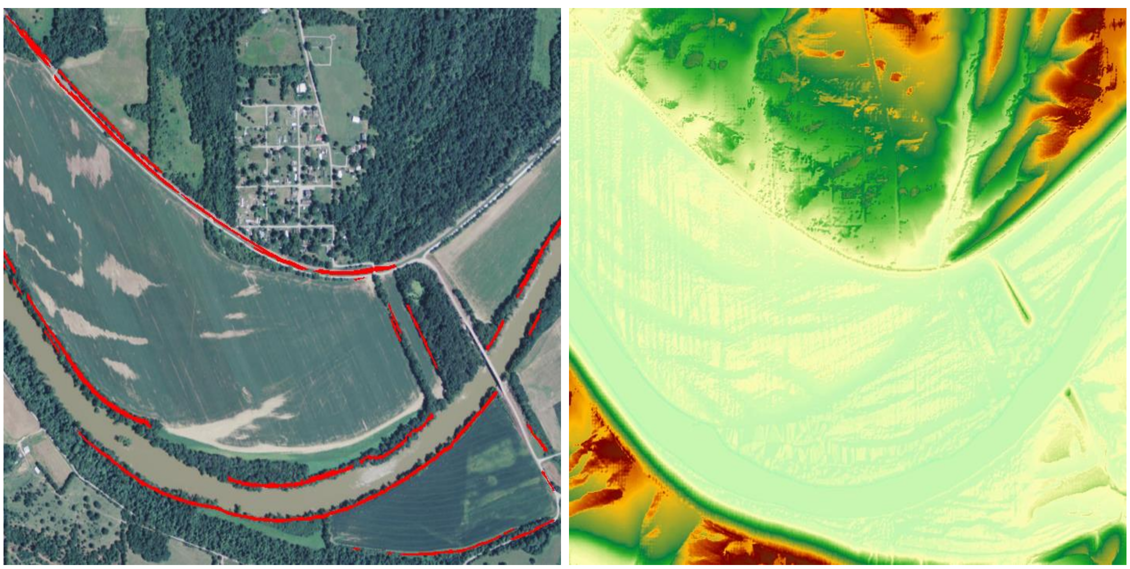


Figure 15c: *Lawrence, before filter (left) and after filter (right) at scale 1:250,000.*



Scale 1: 10,000



Scale 1:10,000

Elevation (ft.)
High : 941.09
Low : 455.03

Figure 15d: *Lawrence County – filtered levee-like structures (left), and symbolized DEM (right).*

DISCUSSION

The purpose of this study is to detect and filter levee-like structures from DEM. The research has two sections, imagery classification, and vector filter. The first section used LiDAR-derived DEM imagery preprocessed by an algorithm that uses Empirical Bayesian Estimation to identify all features significantly above the mean surrounding elevation. The second part aimed to use MBC and Youden Index to locate an optimal threshold value, and use it to differentiate “noise” and levee-like structures. The threshold value was found through a set of survey data in Morgan County, Indiana. The cutoff value was tested in two other Counties. Even though both filtered results proved the efficiency and accuracy of the methods, there were also some factors that affected the results.

Due to the complexity of the landscape, the extracted data usually contains a large amount of “noise”. Unlike engineered levee structures, natural levees can be irregular in the standard and nature of their formation and can be damaged due to water erosion (Courtesy Beeldbank VenW.nl, 2013). The shape of natural levees is greatly influenced by sediment supply and sediment size. Also, flood characteristics, like frequency, duration, seasonality, and magnitude have a considerable impact on the shape of natural levees as well. Therefore, it impossible to preserve all the true positive features from the raw data (Fuhrman, 2000; McGarigal, 2014). For example, the circled features in Figure 16 are natural levees, but they don’t have standard levee characteristics. This kind of small fragments will be removed during the filter process.

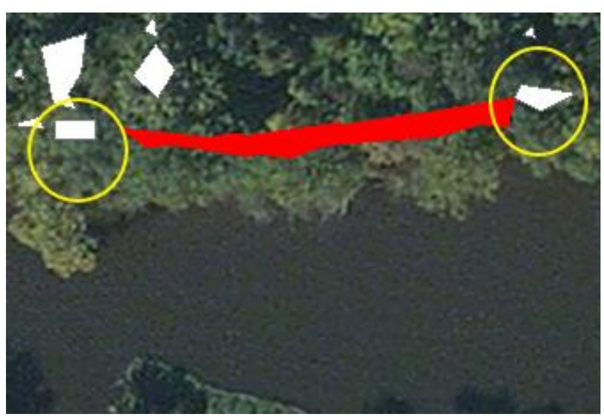


Figure 16: *circled white features are eliminated natural levees, red feature is preserved levee, and white features without circle are eliminated “noise”*

Other than the natural levees, some artificial levees could also be targeted as "noise", such as ring levees (Fuhrman, 2000; Makhdoom, 2013). A ring levee completely encircles or "rings" a small cluster of areas subject to inundation from all directions (Fuhrman, 2000; Makhdoom, 2013), so that the ratio between the feature area and its MBC area is close to 1. 1 minus the high ratio usually falls into the "noise" category, thus this kind of levees will also be removed during the filter process.

There are also other objective reasons that could affect the accuracy of the result, such as (1) the LiDAR sampling rate in the scanning system which may not be dense enough (Bailly et al., 2008); (2) both major and minor outliers are very common due to the vertical error of LiDAR elevation data, as it is not normally distributed (Zandbergen, 2011); (3) the extracted features are quite irregular in shape, such as natural levees; (4) the research lacks of sample data, so the "optimal" threshold value may not be the best one; (5) There exist upward bias in the perimeter lengths' image files as a result of the line segments' stair-stepping pattern. Variation also exists in this bias in terms of its magnitude in relation to the resolution or gain of the image (Figure 17). For these reasons, the area ratio that is computed is bound to be higher than it is in reality (Baker & Cai, 1992; McGarigal, 2014); (6) there are very limited references about filtering vector data of levee.

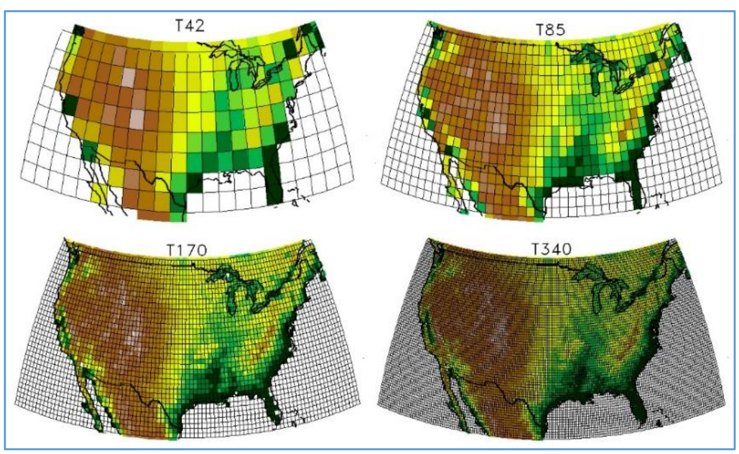


Figure 17: *above pictures are the United States climate model maps in 4 different resolutions. As resolution increases, the grain size decreases, and the proportional abundance of cells of the same class increases. Thus, the measured contagion increases. Even though good resolution reduces the deviation, the sum of grid sizes of the element is still bigger than the actual shape size.*

CONCLUSION

In the first section of this study, elevation was the only determination to extract levee-like structures from DEM. In future studies, the slope can also be taken into account because levees are gently sloped features (Adams et al., 2004). Even though the current classification method is capable of detecting elevated morphological features that have significant height, some levee-like structures can still not be consistently extracted. Future improvement can be focused on referencing more algorithms to the imagery extraction, such as tracking and linking ridge and ravine points (Kweon & Kanade, 1994). This may allow more consistent feature extraction by the program.

The second part of this study attempted to filter the DEM-extracted result. The method uses survey data to locate an optimal threshold ratio to discriminate levee-like structures and “noise”. The amount of sample data, and the places where the data is taken affected the result because having a good amount of sample data can increase precision during the estimation (Ruopp et al., 2008). The surveyed scales were very small, and data quality was not good either. Many levee-like features were not detected in the survey area. Therefore, the calculated results might be biased. By having survey data in more places and larger scale, it will allow the program to compare more threshold values, thus find an even better threshold ratio.

REFERENCE

Adams, Peter N, Slingerland, Rudy L, & Smith, Norman D. (2004). Variations in natural levee morphology in anastomosed channel floodplain complexes. *Geomorphology, 61*(1), 127-142. Retrieved from http://geomorph.geology.ufl.edu/adamsp/Outgoing/Pubs/Adams_EtAl_2004_Geomorph.pdf

Bailey, Mark. (2007). The wonderful world of ... Levees. from http://www.sln.org.uk/geography/schools/blythebridge/6th%20Form%20Work/Levees.ppt.

Bailly, J. S., Lagacherie, P., Millier, C., Puech, C., & Kosuth, P. (2008). Agrarian landscapes linear features detection from LiDAR: application to artificial drainage networks. *International Journal of Remote Sensing, 29*(12), 3489-3508. doi: 10.1080/01431160701469057

Baker, William L, & Cai, Yunming. (1992). The r. le programs for multiscale analysis of landscape structure using the GRASS geographical information system. *Landscape Ecology, 7*(4), 291-302. Retrieved from http://solimserver.geography.wisc.edu/axing/teaching/geog579/lectures/references/Baker_Cai.PDF

Courtesy Beeldbank VenW.nl, Rijkswaterstaat. (2013). *The International Levee Handbook - Chapter 1* (pp. 1350).

Dobigeon, Nicolas, Tourneret, Jean-Yves, & Chang, Chein-I. (2008). Semi-supervised linear spectral unmixing using a hierarchical Bayesian model for hyperspectral imagery. *IEEE Transactions on Signal Processing, 56*(7), 2684-2695. Retrieved from http://oatao.univ-toulouse.fr/803/1/dobigeon_803.pdf

Fawcett, Tom. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*(8), 861-874. Retrieved from http://people.inf.elte.hu/kiss/13dwhdm/roc.pdf

Fluss, Ronen, Faraggi, David, & Reiser, Benjamin. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal, 47*(4), 458-472. Retrieved from https://www.researchgate.net/profile/Benjamin_Reiser2/publication/7600740_Fluss_R_Faraggi_D_Reiser_BEstimation_of_the_Youden_Index_and_its_associated_cutoff_point_Biom_J_47_458-472/links/0912f50f46c4a42455000000/Fluss-R-Faraggi-D-Reiser-BEstimation-of-the-Youden-Index-and-its-associated-cutoff-point-Biom-J-47-458-472.pdf

Fuhrman, Russell L. (2000). Design and Construction of Levees. 1-77. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA402946

Gritzner, Mandy Lineback, Marcus, W Andrew, Aspinall, Richard, & Custer, Stephan G. (2001). Assessing landslide potential using GIS, soil wetness modeling and topographic attributes, Payette River, Idaho. *Geomorphology, 37*(1), 149-165. Retrieved from ftp://130.237.63.22/Common/UllaM/For_Imran/[2001]%20lineback-gritzner%20-%20assessing%20landslide%20potential%20using%20GIS,%20soil%20wetness%20modeling%20and%20topographic%20attributes,%20paye.pdf

Hong, K. (2016). MATLAB TUTORIAL : DIGITAL IMAGE PROCESSING 6 - SMOOTHING : LOW PASS FILTER. Retrieved from http://www.bogotobogo.com/Matlab/Matlab_Tutorial_Digital_Image_Processing_6_Filter_Smoothing_Low_Pass_fspecial_filter2.php

Hudson, Paul F. (2005). Natural levees (Vol. 10): Taylor and Francis. DOI.

Kweon, In So, & Kanade, Takeo. (1994). Extracting topographic terrain features from elevation maps. *CVGIP: image understanding, 59*(2), 171-182. Retrieved from http://riwebbackend.ri.cmu.edu/pub_files/pub3/kweon_in_so_1994_1/kweon_in_so_1994_1.pdf

Li, Zhilin, Zhu, Christopher, & Gold, Chris. (2010). *Digital terrain modeling: principles and methodology*: CRC press.

Makhdoom, M. Tariq. (2013). Impacts of Levees on Floodplain Ecosystems. from http://vaflood.files.wordpress.com/2012/09/levee-impacts-on-floodplain-ecosystems.pdf

McGarigal, K. (2014). Fragstats help, Version 4. *University of Massachusetts*. Retrieved from http://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf

Messner, Frank, & Meyer, Volker. (2006). Flood damage, vulnerability and risk perception–challenges for flood damage research *Flood risk management: hazards, vulnerability and mitigation measures* (pp. 149-167): Springer.

O'Callaghan, John F, & Mark, David M. (1984). The extraction of drainage networks from digital elevation data. *Computer vision, graphics, and image processing, 28*(3), 323-344. Retrieved from ftp://130.237.63.22/Common/UllaM/For_Imran/[1984]%20o'callaghan%20-%20the%20extraction%20of%20drainage%20networks%20from%20digital%20elevation%20data.pdf

Olson, Kenneth R, & Morton, Lois Wright. (2013). Impacts of 2011 Len Small levee breach on private and public Illinois lands. *Journal of Soil and Water Conservation, 68*(4), 89A-95A.

Peucker, Thomas K, & Douglas, David H. (1975). Detection of surface-specific points by local parallel processing of discrete terrain elevation data. *Computer Graphics and image processing, 4*(4), 375-387.

Powers, D.M.W. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and correlation. *Journal of Machine Learning Technologies, 2(1), 37-63.* Retrieved from http://dspace2.flinders.edu.au/xmlui/bitstream/handle/2328/27165/Powers%20Evaluation.pdf?sequence=1

Ruopp, Marcus D, Perkins, Neil J, Whitcomb, Brian W, & Schisterman, Enrique F. (2008). Youden Index and Optimal Cut‐Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical Journal, 50*(3), 419-430. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2515362/

Schisterman, Enrique F, Perkins, Neil J, Liu, Aiyi, & Bondell, Howard. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology, 16*(1), 73-81. Retrieved from http://www4.stat.ncsu.edu/~hdbondel/OptimCut.pdf

Simm, JD, Wallis, Michael, Smith, Philip, Deniaud, Y, Tourment, R, Veylon, G, . . . Glerum, J. (2012). The significance of failure modes in the design and management of levees-a perspective from the International Levee Handbook team. Retrieved from http://eprints.hrwallingford.co.uk/573/1/HRPP539_The_significance_of_failure_modes_in_the_design_and_management_of_levees.pdf

Song, Jeong-Heon, Han, Soo-Hee, Yu, KY, & Kim, Yong-Il. (2002). Assessing the possibility of land-cover classification using lidar intensity data. *International archives of photogrammetry remote sensing and spatial information sciences, 34*(3/B), 259-262. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.4122&rep=rep1&type=pdf

Steinfeld, Celine MM, Kingsford, Richard T, & Laffan, Shawn W. (2013). Semi‐automated GIS techniques for detecting floodplain earthworks. *Hydrological Processes, 27*(4), 579-591.

Tarboton, David G, Bras, Rafael L, & Rodriguez‐Iturbe, Ignacio. (1991). On the extraction of channel networks from digital elevation data. *Hydrological* *processes, 5*(1), 81-100. Retrieved from On the extraction of channel networks from digital elevation data

Tingsanchali, T. (2012). Urban flood disaster management. *Procedia engineering, 32*, 25-37. Retrieved from http://ac.els-cdn.com/S1877705812012647/1-s2.0-S1877705812012647-main.pdf?_tid=2d5eca3a-1508-11e7-9389-00000aacb35e&acdnat=1490851232_55ba71702d862b8423ac7963a667c9c8

Wulder, Michael A, Hall, Ronald J, Coops, Nicholas C, & Franklin, Steven E. (2004). High spatial resolution remotely sensed data for ecosystem characterization. *BioScience, 54*(6), 511-521. Retrieved from http://bioscience.oxfordjournals.org/content/54/6/511.full

Zandbergen, Paul A. (2011). Characterizing the error distribution of lidar elevation data for North Carolina. *International journal of remote sensing, 32*(2), 409-430.

CURRICULUM VITAE

PENG JIN

Education

Bachelor of Science in Computer Science & Social Science – Geography

Troy University, Troy, AL 2007 – 2011

Master of GIS

Indiana University, Indianapolis, IN 2012 – 2017

Honors, Awards, Fellowships

Recipient of ISCP America Scholarship

Recipient of Troy International Scholarship

Recipient of IUPUI fellowship

Research and Training Experience

Network Analysis, Satellite Image Analysis, Linear Referencing, Spatial Statistics, 3D Analysis, VBA, PHP, HTML, SQL, Python

Professional Experience

GIS Developer at INDOT, May 2013 – August 2013

GIS Analyst at CenterPoint Energy, August 2014 – December 2014

GIS Analyst at Kinder Morgan, December 2014 – August 2015

GIS Analyst at UniversalPegasus International, August 2015 – Current